

Intro:

$(x_i, y_i)_{i=1 \dots n}$  dataset connu  
 ↑  
 ex: label f regression  
 feature  $\in \{0,1\}$  classification

But: Trouver un modèle qui explique  $y_i$  en fonction de  $x_i$

Ex: modèle linéaire:

On cherche une relation du type  $y_i \approx w^T x_i = \sum_{k=1}^d w_k x_{ik}$

on cherche  $w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n (y_i - w^T x_i)^2$   
 $= \|y - Xw\|^2$  où  $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ ,  $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$

$$\partial = \nabla_w \|y - Xw\|^2 = -2X^T(y - Xw^*)$$

$$X^T w^* = X^T y$$

$$w^* = (X^T X)^{-1} X^T y$$

least square

Ex: régression logistique:

$$y_i \in \{0,1\}$$

modèle:  $y_i$  suit une Bernoulli de paramètre  $g(x_i^T w)$

$$P_{w^*}(y_i=1 | x_i) = g(x_i^T w)$$

$$P_{w^*}(y_i=0 | x_i) = 1 - g(x_i^T w)$$

$$\text{Vraisemblance: } P_w(y_1=y_1, \dots, y_n=y_n | x_1, \dots, x_n) = \prod_{i=1}^n P_w(y_i=y_i | x_i) = \prod_{i=1}^n g(x_i^T w)^{y_i} (1-g(x_i^T w))^{1-y_i}$$

$$\text{Estimateur du max de vraisemblance: } w^* = \underset{w}{\operatorname{argmax}} \sum_{i=1}^n g(x_i^T w)^{y_i} (1-g(x_i^T w))^{1-y_i}$$

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n -y_i \ln g(x_i^T w) - (1-y_i) \ln (1-g(x_i^T w))$$

Pas d'expression explicite ...

Approcher  $w^*$  par un algorithme itératif

Cas plus générale: minimisation du risque empirique

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n L(w, x_i, y_i) + R(w) \quad L = \text{fonction de perte (loss)}$$

$$\text{ex: } L(w, x_i, y_i) = (y - x^T w)^2 \rightarrow \text{régression linéaire}$$

$$L(w, x_i, y_i) = -y_i \ln g(x^T w) - (1-y_i) \ln (1-g(x^T w)) \rightarrow \text{régression logistique}$$

$$L(w, x_i, y_i) = \text{hinge loss} \rightarrow \text{Support Vector Machines}$$

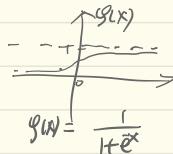
$$R(w) = \text{regularisation} \quad \text{ex: } R(w) = \lambda \|w\|^2 \rightarrow \text{ridge, ribbonov}$$

$$R(w) = \lambda \|w\|_1 \rightarrow \text{LASSO}$$

$$\|w\|_1 = \sum_{k=1}^d |w_k|$$

$$w^* = \underset{w}{\operatorname{argmin}} \|y - Xw\|^2 + \lambda \|w\|^2 \quad \theta = -X^T(y - Xw^*) + \lambda w^* = -X^T y + (\lambda I + X^T X)w^*$$

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$



# Éléments d'analyse convexe

Posons  $X = \mathbb{R}^d$

## I. Ensembles convexes, fonctions convexes

$C \subset X$

Def:  $C$  est convexe si



$$\forall x, y \in C, \quad \forall t \in [0, 1], \quad tx + (1-t)y \in C$$



Soit  $f: X \rightarrow [-\infty, +\infty]$

[valeur très importante] Fonction indicatrice d'un ensemble  $C$

$$c(x) = \begin{cases} 0 & \text{pour } x \in C \\ +\infty & \text{pour } x \notin C \end{cases}$$

)

domaine  $\text{dom}(f) = \{x \in X : f(x) < +\infty\}$  en:  $\text{dom}(c_0) = C$

épi graphe:  $\text{epi}(f) = \{(x^*, y) \in X \times \mathbb{R} : y \geq f(x^*)\}$



Def:  $f: X \rightarrow [-\infty, +\infty]$  est convexe si  $\text{epi}(f)$  est convexe

propriété:  $f$  est convexe ssi  $\forall x, y \in \text{dom}(f), \forall t \in [0, 1], \quad f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$

preuve:  $\Rightarrow$  Supposons  $f$  convexe

$$(x, f(x)) \in \text{epi}(f), \quad (y, f(y)) \in \text{epi}(f)$$

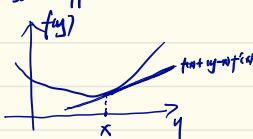
$$- \forall t \in [0, 1], \quad t(x, f(x)) + (1-t)(y, f(y)) \in \text{epi}(f)$$

$$(tx + (1-t)y, tf(x) + (1-t)f(y)) \in \text{epi}(f)$$

$$\Leftrightarrow t(x, f(x)) + (1-t)(y, f(y)) \ni tx + (1-t)y$$



## II. Somme différentielle



toute fonction convexe dérivable domine sa tangente  
en  $x$

$$f(y, f(y)) \geq f(x) + c(x), \quad y \rightarrow x$$

preuve:  $d(x)$  est le vecteur tel que:

$$f(x+th) = f(x) + c(x, h) + o(th)$$

$$h = t(y-x)$$

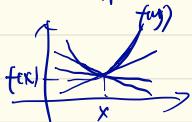
$$f(x+th) = f(x) + t(c(x, y-x) + o(t))$$

$$f(x+ty-x) = f((1-t)x+ty) \leq (1-t)f(x) + t f(y)$$

$$\text{Donc } (1-t)f(x) + t f(y) \geq f(x) + t f(y), \quad y-x > t(x)$$

$$\text{on fait } t \rightarrow 0 \quad f(y) - f(x) \geq \langle \nabla f(x), y-x \rangle$$

$f$  non différentiable :



$\nabla f(x) = \text{ensemble des pôles des "tangences"}$

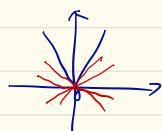
Déf :  $g \in \mathbb{R}$  est un sous-gradient de  $f$  au point  $x \in \text{dom } f$  si

$$\forall y, \quad f(y) \geq f(x) + \langle g, y-x \rangle$$

$$\nabla f(x) = \{g \in \mathbb{R}^n \mid g \text{ est un sous-gradient en point } x\}$$

on pose  $\nabla f(x) = \emptyset$  si  $x \notin \text{dom } f$ , par convention

$$\text{Ex: } f(x) = |x| \quad (x \in \mathbb{R})$$



$$\nabla f(x) = \begin{cases} \{y\} & \text{si } x > 0 \\ \{-y\} & \text{si } x < 0 \\ [-1, 1] & \text{si } x = 0 \end{cases}$$

$$\begin{aligned} y &\in \nabla f(0) \Leftrightarrow \forall y \in \mathbb{R}, \quad |y| \geq y \cdot y \\ &\Leftrightarrow y \geq 0, \quad \frac{|y|}{y} \geq 1 \\ &y \leq 0, \quad \frac{|y|}{y} \leq -1 \\ &\Leftrightarrow -1 \leq y \leq 1 \end{aligned}$$

Propriété : si  $f$  est dérivable au point  $x$ , alors  $\nabla f(x) = \{f'(x)\}$

Preuve :

Propriété (admise) : Soit  $f: X \rightarrow [-\infty, +\infty]$

Mais :  $\forall x \in \text{int}(\text{dom}(f)), \quad \nabla f(x) \neq \emptyset$       int = intérieur



$$\nabla f(x) \subseteq \partial f(x)$$

soit  $y \in \nabla f(x)$ , Montrons que  $y = \nabla f(x)$

$$f(x+h) \geq f(x) + \nabla f(x), h > 0$$

$$f(x+t(y-\nabla f(x))) \geq f(x) + t(y-\nabla f(x)) + \nabla f(x)$$

$$\text{on a } f(x+t(y-\nabla f(x))) \geq f(x) + t(y-\nabla f(x)) + \nabla f(x) - t \nabla f(x)$$

$$= f(x) + t \langle y, y - \nabla f(x) \rangle$$

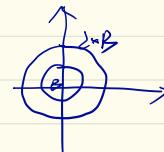
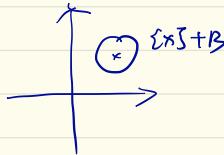
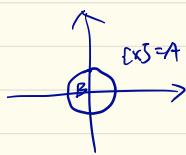
$$f(x) + t \langle y, y - \nabla f(x) \rangle \leq f(x) + t \langle \nabla f(x), y - \nabla f(x) \rangle + t \nabla f(x)$$

$$t > 0 \quad \langle y, y - \nabla f(x) \rangle \leq \langle \nabla f(x), y - \nabla f(x) \rangle$$

$$\|y - \nabla f(x)\|^2 \leq 0 \quad \text{donc} \quad y = \nabla f(x)$$

$f: X \rightarrow P(X)$

Remarque :  $A+B = \{a+b : a \in A, b \in B\}$   
 $M \cdot B = \{M \cdot B^j : b \in B\}$  où  $M$  est une matrice



Règles de calcul:  $x \in \mathbb{R}^d$ ,  $y \in \mathbb{R}^m$

$f: X \rightarrow [-\infty, +\infty]$     $g: Y \rightarrow [-\infty, +\infty]$     $M: X \rightarrow Y$  op. linéaire (=matrix mod.)

$(f+g)M: x \mapsto f(x)+g(Mx)$

$\triangleright f, g$  différentiable:  $D(f+g)M(x) = Df(x) + Dg(Mx) = Df(x) + M^T Dg(Mx)$

$Df(x) + M^T Dg(Mx) \subset D(f+g)M(x)$

Preuve:  $Df(x) + M^T Dg(Mx) = \{g + g' : g \in Df(x), g' \in M^T Dg(Mx)\}$

$= \{g + M^T \psi : g \in Df(x), \psi \in Dg(Mx)\}$

Soit  $g \in Df(x)$ ,  $\psi \in Dg(Mx)$ ,  $\forall y, f(y) \geq f(x) + \langle g, y-x \rangle$ ,  $g(My) \geq g(Mx) + \langle \psi, My-Mx \rangle$  vrai pour  $j=Mg$

$\forall y, f(y) + g(My) \geq f(x) + g(Mx) + \langle g, y-x \rangle + \langle \psi, My-Mx \rangle$

$\forall y, (f+g)M(y) \geq (f+g)M(x) + \langle g + M^T \psi, y-x \rangle$

Donc  $g + M^T \psi \in D(f+g)M(x)$

Résultat

Si  $\sigma \in \text{int}(M\text{dom}f - \text{dom}g)$ , alors  $(f+g)M(\sigma) = f(\sigma) + M^T Dg(\sigma)$

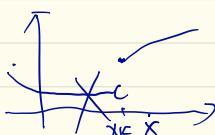
### III: Minimiseurs

s.c.i

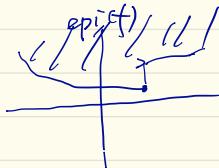
Déf:  $f: X \rightarrow [-\infty, +\infty]$  est dite semi-contINUE inférieurement en un point  $x$

si  $\forall (x_n) \xrightarrow{n \rightarrow +\infty} x$ ,  $\liminf_{n \rightarrow +\infty} f(x_n) \geq f(x)$

Rappel:  $(\liminf f(x_n)) = \lim_{n \rightarrow +\infty} \inf_{k \geq n} f(x_k)$



Def:  $f$  est s.ci si elle est s.ci en tout point  $x$



propriété:  $f$  s.ci  $\Leftrightarrow \text{epi}(f)$  est fermé

Def:  $f$  est dite coercive si  $f(x) \xrightarrow{\|x\| \rightarrow \infty} \infty$

propriété: si  $f$  est sci coercive,  $\text{argmin} f \neq \emptyset$

Preuve: On, il existe  $m$  tel que  $f(m) \leq \inf f + \frac{1}{n}$   $\forall x^* \text{ tel que } f(x^*) = \inf f(x)$   
 $f(x_n) \xrightarrow{n \rightarrow \infty} \inf f(x^*)$



Soit  $x^*$  une valeur d'adhérence de  $x_n$ , c'est à dire  $\exists q_n, x_{q_n} \rightarrow x^*$

Il existe car  $(x_n)$  est bornée!

Par l'hypothèse, si  $x_n$  n'était pas bornée, il existerait  $q_n$  tel que  $\|x_{q_n}\| \rightarrow \infty$

Alors  $f(x_{q_n}) \rightarrow \infty$  car  $f$  coercive contredit C\*)

Comme  $f$  est sci,  $\inf f(x_{q_n}) \geq f(x^*)$

$\inf f \geq f(x^*)$  donc  $x^* \in \text{argmin} f$

Def:  $f: X \rightarrow [-\infty, \infty]$  est dite strictement convexe si  $\forall x, y \in \text{dom} f, \forall t \in ]0, 1[$ ,

$$f(tx + (1-t)y) < t f(x) + (1-t) f(y)$$

Ex:  $f(x) = x^2 \quad f(x) = e^x$

propriété:  $f$  strictement convexe admet au plus un minimiseur

Preuve: soient  $x^* \neq y^*$  deux minimiseurs

$$\begin{aligned} f\left(\frac{x^* + y^*}{2}\right) &< \frac{1}{2} f(x^*) + \frac{1}{2} f(y^*) = \frac{1}{2} \inf f + \frac{1}{2} \inf f \\ &< \inf f \quad \text{Absurde} \end{aligned}$$

Def:  $f: X \rightarrow [-\infty, \infty]$  est dite  $\mu$ -familièrement convexe ( $\mu > 0$ ) si

$$x \mapsto f(x) - \frac{\mu}{2} \|x\|^2 \text{ est convexe}$$



Propriété :  $f$  fortement convexe  $\Rightarrow f$  coercive  
 $f$  —  $\Rightarrow f$  strictement croissante

Corollaire :  $f$  sci fortement convexe admet un unique minimiseur

## Chapitre 3

### Algorithme du gradient proximal

#### I. opérateur proximal

Soit  $P_0(x) = \text{ensemble des fonctions convexes, sci. de domaine } f \neq \emptyset$   
 Def : Soit  $g \in P_0(x)$ ,  $\forall x \in X$

$$\text{prox}_g(x) = \arg \min_y g(y) + \frac{\|y-x\|^2}{2}$$

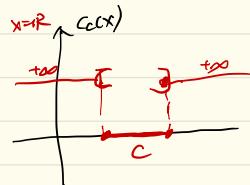
l'argmin est un singleton car  $y \mapsto g(y) + \frac{\|y-x\|^2}{2}$  est 1-fortement croissant

Ex : ①  $g(y) = \ell_c(x)$  où  $c \in C_X$ ,  $c$  convexe, fermé, non vide

$$\text{prox}_{\ell_c}(x) = \arg \min_y \ell_c(y) + \frac{\|y-x\|^2}{2}$$

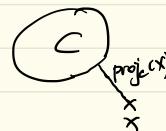
$$= \arg \min_{y \in C} (\ell_c(y) + \frac{\|y-x\|^2}{2}) = \arg \min_{y \in C} \|y-x\|$$

$$= \text{le point de } C \text{ le plus proche de } x = \text{proj}_C(x)$$



$$\textcircled{2} \quad g(x) = |x|, \quad x \in \mathbb{R}$$

$$\text{prox}_g(x) = \arg \min_{y \in \mathbb{R}} (|y| + \frac{(y-x)^2}{2})$$



rule de Fermat :

$$p = \text{prox}_{\ell_c}(x) \text{ satisfait : } 0 \in \nabla(\ell_c(\cdot) + \frac{\| \cdot - x \|^2}{2})(p)$$

$$\nabla \ell_c(\cdot) + \frac{\| \cdot - x \|^2}{2}(p) = \nabla(\ell_c(\cdot))(p) + \nabla(\frac{\| \cdot - x \|^2}{2})(p) \quad y \mapsto \frac{\|y-x\|^2}{2} = \text{sign}(p) + p \cdot x$$

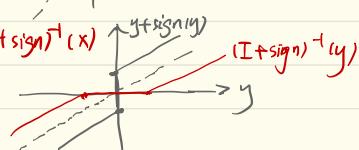
$$\text{or } \text{sign}(p) = \begin{cases} 1 & \text{si } p > 0 \\ -1 & \text{si } p < 0 \\ 0 & \text{si } p = 0 \end{cases}$$

$$p = \text{prox}_{|\cdot|}(x) \Leftrightarrow 0 \in \text{sign}(p) + p \cdot x$$

$$\Leftrightarrow x \in p + \text{sign}(p)$$



$$\Leftrightarrow p = (I + \text{sign})^{-1}(x)$$



## Fonction de sautage deux

$$S_0(x) = \begin{cases} 0 & \text{si } |x| \leq r \\ x-r & \text{si } x > r \\ x+r & \text{si } x < -r \end{cases}$$

$$\begin{cases} \text{prox}_{S_0}(x) = S_0(x) \\ \text{prox}_{S_0^*}(x) = S_0(x) \end{cases}$$

## II Algorithme du gradient proximal

$\min_{x \in X} f(x) + g(x)$  où  $f$  est différentiable,  $g$  bornée

$$x^{k+1} = \text{prox}_g^{-1}(x^k - \gamma \nabla f(x^k)) = \arg \min_{y \in X} (y \cdot g(y) + \frac{\gamma \|y - x^k\|^2}{2})$$

11.16

Rappels :

$$f: X \rightarrow (-\infty, +\infty]$$

$$\text{dom } f = \{x : f(x) < \infty\}$$

$$\nabla f(x) = \{v : v_j, f'_j \geq f(x) + \langle v_j, \cdot - x \rangle \quad (\forall j)\}$$

$$\nabla^2 f(x) = \nabla^T \nabla f(x) = \nabla(\nabla f(x))^T \quad \text{si } \exists \Omega \in \text{int}(\text{dom } f - \text{dom } g)$$

(ou plus généralement,  
intérieur relatif\*)

[Algorithm du gradient :

$$\min_{x \in X} f(x) \quad x^{k+1} = x^k - \gamma \nabla f(x^k)$$

Application :  $f(x) = \text{terme d'attache aux données}$   
 $= \frac{1}{n} \sum_{i=1}^n l(x_i, y_i)$

$g(x) := \text{terme de régularisation}$

$$= \begin{cases} \frac{\lambda}{2} \|x\|^2 & (\text{ridge}) \\ \lambda \|x\|_1 & (\text{lasso}) \end{cases}$$

(c où  $C$  convexe fermé non vide  
(contrainte  $x \in X$ )

$$x^{k+1}, \|x^{k+1}\|_1 = \lambda \sum_{i=1}^n |x_i|$$

Calculons  $\text{prox}_{S_0^*}(x)$

$$p = \text{prox}_{S_0^*}(x) \Leftrightarrow p = \arg \min_p (p \cdot y + \frac{\gamma \|p\|^2}{2}) \Leftrightarrow 0 \in \partial(-\gamma p) \quad (\text{Règle de Fermat}) \Leftrightarrow 0 \in \gamma \text{sign}(p) + p - x$$

$$\Leftrightarrow x \in p + \gamma \text{sign}(p) \Leftrightarrow x = p + \gamma, p > 0 \Leftrightarrow p = x + \gamma, x \geq 0$$

$$\text{ou } x = p - \gamma, p < 0 \Leftrightarrow p = x - \gamma, x \leq 0$$

$$\text{ou } x \in p + \gamma \mathbb{I}_{\{1\}}, p = 0 \Leftrightarrow p = 0, x \in [x, x + \gamma]$$

$$\Leftrightarrow \text{prox}_{S_0^*}(x) = S_\gamma(x)$$

Calculons  $\text{prox}_{\lambda \|x\|_1}(x)$

$$\min_{y \in X} f(y) + f(x) \quad x^* = \arg \min_x f(x), x^* = \arg \min_y f(y)$$

$$p^* = \text{prox}_{\lambda \|x\|_1}(x) \Leftrightarrow p^* = \arg \min_{y \in X} \sum_{i=1}^n (y_i - x_i)^2 + \frac{\lambda \|y\|^2}{2} \Leftrightarrow p^* = \arg \min_{y \in X} \sum_{i=1}^n (y_i - x_i)^2 + \frac{\lambda \|y\|^2}{2} \quad \Leftrightarrow \text{Bi}, p^* = \arg \min_{y \in X} (\lambda \|y\| + \frac{\lambda \|y\|^2}{2})$$

$$\Leftrightarrow \text{prox}_{\lambda \|x\|_1}(x) = (S_\lambda(x), S_\lambda(x), \dots, S_\lambda(x))$$

Algorithm ISTA : (Interactive soft-threshold)

Pb :  $x \in \mathbb{R}^n$  vecteur de features,  $y \in \mathbb{R}$  réponse / label

On veut chercher un modèle du type:  $y_i \approx x_i^T x$

= LASSO!  $\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N (y_i - x_i^T x)^2 + \lambda \|x\|_1 \Leftrightarrow \min_{x \in \mathbb{R}^n} \frac{1}{N} \|Ax - \bar{y}\|^2 + \lambda \|x\|_1$  où  $\bar{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$  et  $A = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix}$

Algorithm:

$$x^{k+1} = \text{prox}_{\lambda \| \cdot \|_1} (x^k - \frac{1}{N} A^T (A x^k - \bar{y}))$$

↑  
seuillage dans de chaque composante

Autre exemple:  $\min_{x \in \mathbb{R}^n} f(x) \Leftrightarrow \min_{x \in \mathbb{R}^n} f(x) + C(x)$

$$x^{k+1} = \boxed{\text{prox}_{C(x)}}(x^k - \nabla f(x^k))$$

$\text{proj}_C = \text{projection}$

Justification:  $x^{k+1} = \text{prox}_{Cg}(x^k - \nabla f(x^k))$

Algorithm du point fixe:  $x^{k+1} = T(x^k)$  où  $T$  est continue

Si  $x^k$  converge vers  $x^*$  alors  $x^* \in \arg\min f + g$

Preuve:  $x^*$  est un point fixe de  $T$ ,  $x^* = T(x^*)$ .

$$x^* = \text{prox}_{Cg}(x^* - \nabla f(x^*)) \Leftrightarrow x^* = \arg\min_y (f(y) + \frac{\|y - (x^* - \nabla f(x^*))\|^2}{2})$$

$$\Leftrightarrow \exists \varepsilon \in \mathbb{R} \quad \forall y \in \mathbb{R} \quad f(y) + \frac{\|y - (x^* - \nabla f(x^*))\|^2}{2} \geq f(x^*) + \frac{\|(x^* - \nabla f(x^*)) - (x^* - \nabla f(x^*))\|^2}{2} \Leftrightarrow x^* \in \arg\min f + g$$

$$x^* = T(x^*) \text{ équivalent à: } \exists \varepsilon \in \mathbb{R} \quad f(x^*) + \frac{\|x^* - (x^* - \nabla f(x^*))\|^2}{2} \leq f(x^*) + \frac{\|(x^* - \nabla f(x^*)) - (x^* - \nabla f(x^*))\|^2}{2} \Leftrightarrow x^* \in \arg\min f + g$$

$$\Leftrightarrow x^* \in \arg\min f + g$$

Algorithm du gradient stochastique:

SGD = stochastic gradient descent

Pb:  $\min_{x \in X} F(x)$  où  $F(x) = \mathbb{E}(f(x, \xi))$  où  $\xi: \Omega \rightarrow \mathbb{R}$  est une variable aléatoire  
 $f(\cdot, \xi)$  est convexe, dérivable

Sousst: On n'est pas capable de calculer l'espérance, car on ne connaît pas la loi de  $\xi$   
ou on la connaît, mais le calcul de l'intégrale  $\mathbb{E}$  est coûteux

En revanche, on observe des réalisations i.i.d de  $\xi$ .

$\xi_1, \xi_2, \xi_3, \dots$  iid même loi que  $\xi$

Algorithme:  $\boxed{x^{k+1} = x^k - \nabla f(x^k, \xi_{k+1})}$  où  $\nabla f(x, \xi)$  est la dérivée par rapport à  $x$  de  $f(x, \xi)$

$x^{k+1} = x^k - \underbrace{\nabla \mathbb{E}_{\xi \sim P}(f(x^k, \xi))}_{F(x^k)}$  Gradient Descent

Initialisation  $x^0$ : on observe la fonction  $f(\cdot, \xi_1)$

Exemple : Minimisation du risque empirique

$$\min_{x^k} \frac{1}{N} \sum_{i=1}^N L(x_i, (x_i, y_i))$$

A l'instant  $k$ , on a  $x^k \rightarrow$  on choisit défaillamment  $I_{k+1} \in \{1, \dots, N\}$

$$\rightarrow \text{On actualise : } \boxed{x^{k+1} = x^k + Df(x^k, (x_{I_{k+1}}, y_{I_{k+1}}))}$$

$$(x_i, y_i), \quad - \quad (x_{I_{k+1}}, y_{I_{k+1}})$$

Dans cet exemple :  $\varepsilon_{k+1} = I_{k+1} = \text{l'indice}$

$$f(x^*) = L(x^*, (x_i, y_i))$$

On minimise alors la fonction  $F(x) = E_\varepsilon(L(x, (x_i, y_i))) = E(L(x, (x_\varepsilon, y_\varepsilon))) = \frac{1}{N} \sum_{i=1}^N L(x, (x_\varepsilon, y_i)) = \frac{1}{N} \sum_{i=1}^N L(x, (x_i, y_i))$

Analyse : soit  $x^*$  un minimiseur de  $F(x) = E_\varepsilon(L(x, \varepsilon))$  (on suppose qu'il en existe)

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^* + Df(x^k, \varepsilon_{k+1})\|^2 = \|x^k - x^*\|^2 + 2\langle x^k - x^*, Df(x^k, \varepsilon_{k+1}) \rangle, \quad x^k - x^* + Df(x^k, \varepsilon_{k+1}) \parallel$$

$$\Rightarrow \|x^{k+1} - x^*\|^2 |_{\varepsilon_1 = k+1} \leq \|x^k - x^*\|^2 - 2\langle Df(x^k, \varepsilon_{k+1}) |_{\varepsilon_1 = k+1}, x^k - x^* \rangle, \quad x^k - x^* + Df(x^k, \varepsilon_{k+1})$$

$$\leq \|x^k - x^*\|^2 - 2\langle Df(x^k), x^k - x^* \rangle$$

Hypothèse :  $\forall k, \mathbb{E}[ \|Df(x, \varepsilon)\|^2] \leq C$

$$\mathbb{E}[Df(x, \varepsilon)] = Df(x, \varepsilon) = Df(x)$$

Convexité de  $F$  :  $F(x^*) \geq F(x^k) + \langle \nabla F(x^k), x^* - x^k \rangle$

$$f_k(F(x^k) - F(x^*)) \leq \langle x^k - x^*, x^* - x^k \rangle \leq \frac{\|x^k - x^*\|^2}{2} - \mathbb{E}\left(\frac{\|x^{k+1} - x^*\|^2}{2}\right) (\varepsilon_{k+1}) + C \frac{k^2}{2}$$

$$\begin{cases} f_k(F(x^k) - F(x^*)) \leq \mathbb{E}\left(\frac{\|x^k - x^*\|^2}{2}\right) - \mathbb{E}\left(\frac{\|x^{k+1} - x^*\|^2}{2}\right) + C \frac{k^2}{2} \\ f_{k+1}(F(x^k) - F(x^*)) \leq \mathbb{E}\left(\frac{\|x^k - x^*\|^2}{2}\right) - \mathbb{E}\left(\frac{\|x^{k+1} - x^*\|^2}{2}\right) + C \frac{(k+1)^2}{2} \end{cases}$$

$$\text{Finallement : } \frac{\sum_{k=1}^K f_k(F(x^k) - F(x^*))}{\sum_{k=1}^K f_k} \leq \frac{\|x^k - x^*\|^2 + C \frac{\sum_{k=1}^K k^2}{K}}{\sum_{k=1}^K f_k}$$

$$\frac{\sum_{k=1}^K f_k(F(x^k))}{\sum_{k=1}^K f_k} \geq F\left(\frac{\sum_{k=1}^K x^k}{\sum_{k=1}^K f_k}\right) \quad (\text{Jensen})$$

Puis  $\bar{x}^k = \frac{\sum_{i=1}^k x^i}{\sum_{i=1}^k f_i}$  estime moyenne

$$F(\bar{x}^k) - F(x^*) \leq \frac{\|x^k - x^*\|^2 + C \frac{\sum_{k=1}^K k^2}{K}}{\sum_{k=1}^K f_k}$$

$$\left[ \sum_{k=1}^K k^2 \geq \infty \text{ et } \sum_{k=1}^K f_k = \infty \right]$$

$$f_k = \frac{1}{K}$$

Si on veut que  $F(\bar{x}^k) \rightarrow F(x^*)$ , il suffit que

meilleur choix  $f_k \sim \frac{1}{K}$

## TD 2 SD-TS2A 211

Exercice 9 1)  $\text{IP}(Y_i = -1 | x_i) = 1 - \text{IP}(Y_i = 1 | x_i)$

2)  $\text{IP}(Y_1 = y_1, \dots, Y_n = y_n | x_1, \dots, x_n)$   
 $= \prod_{i=1}^n \text{IP}(Y_i = y_i | x_i)$

$$L = \log \prod_{i=1}^n \text{IP}(Y_i = y_i | x_i) = -\sum_{i=1}^n \log (1 + \exp(-y_i(x_i^T w + w_0)))$$

3)  $\nabla f(w, w_0) = \begin{pmatrix} \nabla_w f(w, w_0) \\ \nabla_{w_0} f(w, w_0) \end{pmatrix} = -\sum_{i=1}^n \frac{y_i \exp(-y_i(x_i^T w + w_0))}{1 + \exp(-y_i(x_i^T w + w_0))} \begin{pmatrix} x_i \\ 1 \end{pmatrix}$

4)  $\underset{g}{\text{prox}}(x) = \arg \min_g [g(y) + \frac{1}{2} \|x - y\|^2]$

$$g(x) = \frac{\lambda}{2} \|x\|^2 \quad \in \partial g(p) + \partial h(p)$$

Règle de Fermat:  $\Leftrightarrow 0 \in \lambda p + x - p \quad \Leftrightarrow p = \frac{x}{\lambda+1}$

5) Trier  $i \in \{1, \dots, n\}$  uniformément

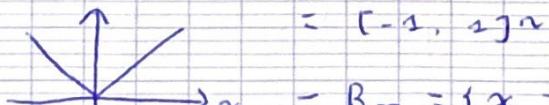
\*  $x_{k+1} = \text{prox}_g(x_k - \gamma \nabla f_i(w, w_0))$

$$\nabla f(w, w_0) = \sum_{i=1}^n \nabla f_i(w, w_0).$$

$$x_{k+1} = \frac{1}{\lambda+1} (x_k - \gamma \nabla f_i(w, w_0))$$

Exercice 7 1)  $0 \in A^T(Ax^* - b) + \underbrace{\partial(\| \cdot \|_2)}_{\geq 0}(x^*)$

$$\partial(\| \cdot \|_2)(z) = (\alpha(1, 1)1_0), \dots, \alpha(1, 1)1_n$$



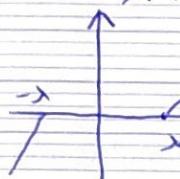
$$B_\infty = \{x \in \mathbb{R}^2 \mid \|x\|_\infty \leq 1\}$$

$$\forall i, \frac{(A^T b)_i}{\lambda} \in [-1, 1] \quad \forall i, |\lambda| \geq |(A^T b)_i|$$

$$\lambda \geq \|A^T b\|_\infty$$

$$2) \alpha_{k+1} = \frac{\text{prox}_{\lambda \|\cdot\|_2}(\alpha_k - \gamma A^T(A\alpha_k - b))}{\lambda \|\cdot\|_2}$$

$$S = \text{prox}_{\lambda \|\cdot\|_2}$$



$$= (S(\alpha_{k,1} - \gamma A^T(A\alpha_{k,2} - b)), \dots, S(\dots))$$

$$3) 0 < \gamma < \frac{2}{L}, \quad f + g \neq \emptyset$$

$$F(x_k) = \inf F \leq \frac{L \|x_0 - x^*\|^2}{2k}$$

$$\frac{LD^2}{2k} \leq \epsilon$$

$$k \geq \frac{LD^2}{2\epsilon}$$

~~Exercice 2~~

Exercice 1:

$$2) \partial L_C(x) = \{g \in \mathbb{R}^n, \forall y \in C,$$

$$\partial f(x) = \{y \in \mathbb{R}^n, \forall y \in \text{dom } f, f(y) \geq f(x) + \langle y - x, y \rangle\}$$

$$\text{Si } x \in C, f(x) = 0.$$

$$x \notin C, C_C(x) = \emptyset$$

$$3). 0 \in \underbrace{\nabla f(x^*)}_{} + \partial L_C(x^*)$$



$$4) \partial L_{H_{w,b}}(x) = \{g \in \mathbb{R}^n, \forall y \in H_{w,b}, \langle g, y - x \rangle \leq 0\}$$

$$\text{Si } x \in H_{w,b} \quad \partial L_{H_{w,b}}(x) = \{x\}$$

$$5) \underset{x}{\operatorname{argmin}} \underbrace{\frac{1}{2} \|x - z\|_2^2}_{f(x)}$$

$$-(x^* - z) \in \lambda w \quad \exists \lambda, -(x^* - z) = \lambda w$$

$$\langle w, x^* \rangle + b = 0 \quad \langle w, -\lambda w + z \rangle + b = 0$$

$$-\lambda \|w\|^2 + \langle w, z \rangle + b = 0. \quad \lambda = \frac{\langle w, z \rangle + b}{\|w\|^2}$$

$$\|x^* - z\|_2 = \|-\lambda w + z - z\|_2 = |\lambda| \|w\|_2$$

$$= \frac{|\langle w, z \rangle + b|}{\|w\|_2}$$

## Dualité lagrangienne

$X = \mathbb{R}^n$      $\min_{x \in C} f(x)$      $\begin{matrix} \text{contrainte d'égalité affine} \\ \text{contrainte d'inégalité} \end{matrix}$

$$C = \{x \in X, Ax = b, g_1(x) \leq 0, \dots, g_p(x) \leq 0\}$$

on  $A \in \mathbb{R}^{n \times m}$      $b \in \mathbb{R}^m$      $g_1, \dots, g_p$  convexe

Soit  $f: X \rightarrow [-\infty, +\infty]$

$g_1, \dots, g_p: X \rightarrow \mathbb{R}$  conv

$A \in \mathbb{R}^{n \times n}$      $b \in \mathbb{R}^m$

Notation  $\vec{a} \leq \vec{0}$  signifie  $\forall i, a_i \leq 0$ .

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_p(x) \end{pmatrix}$$

Pb  $\min f(x)$

$x$      $g(x) \leq 0$  et  $Ax = b$

$$\text{fond } \{g \leq 0, y = \{x \in X, g(x) \leq 0\}\}$$

$$A^{-1}b = \{x \in X; Ax = b\}$$

le problème me revient à minimiser la fonction

$$x \mapsto f(x) + L_{\{g \leq 0\}}(x) + L_{A^{-1}b}(x)$$

l'espace des contraintes  $C = \{g \leq 0\} \cap A^{-1}b$

Fonction primale  $x \mapsto f(x) + L_{\{g \leq 0\}}(x) + L_{A^{-1}b}(x)$

Valeur primaire  $p = \inf_{x \in C} f(x) + L_{\{g \leq 0\}}(x) + L_{A^{-1}b}(x)$

Un solution primaire est un minimisation de la fct primal  
s'il en existe, l'inf est atteint :  $p = \inf_{x \in C} f(x) + \dots$

Def:  $\forall (x, \lambda, v) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$

$$\begin{aligned} L(x, \lambda, v) &= f(x) + \langle \lambda, Ax - b \rangle + \langle v, g(x) \rangle - L_{A^{-1}b}(v) \\ &= -\infty \text{ si } v \text{ n'est pas positif} \end{aligned}$$

Ex cas de contrainte d'égalité seulement ( $p=0$ )

$$L(x, \lambda) = f(x) + \langle \lambda, Ax - b \rangle$$

Lemme:

$$\forall x, f(x) + L_{\{g \leq 0\}}(x) + L_{A^{-1}b}(x) = \sup_{(\lambda, v) \in \mathbb{R}^m \times \mathbb{R}^p} L(x, \lambda, v)$$

Preuve: cas de contrainte d'égalité

$$L(x, \lambda) = f(x) + \langle \lambda, Ax - b \rangle$$

$$\begin{aligned} \sup_x L(x, \lambda) &= \begin{cases} f(x) & \text{si } Ax - b = 0 \\ +\infty & \text{sinon} \end{cases} \\ &= f(x) + L_{A^{-1}b}(x) \end{aligned}$$

$$p = \inf_x \sup_{(\lambda, v)} L(x, \lambda, v)$$

Def La valeur duale est

$$d = \sup_{(\lambda, v)} \inf_x L(x, (\lambda, v))$$

la fonction duale est

$$\bar{L}(\lambda, v) = \inf_x L(x, (\lambda, v))$$

Solution duale = maximiseur de  $\bar{L}$

Proposition  $d \leq p$  (équation faite)

Quand a-t-on  $p = d$  ?

Déf  $(x, \varphi) \in \mathbb{R}^n \times (\mathbb{R}^m \times \mathbb{R}^P)$  est un point-selle de  $\mathcal{L}$

Si  $x \in \arg\min \mathcal{L}(\cdot, \varphi)$

$\varphi \in \arg\max \mathcal{L}(x, \cdot)$

Théorème  $(x, \varphi)$  point-selle de  $\mathcal{L}$  si et seulement si

$$\begin{cases} x \text{ est solution primal} \\ \varphi \text{ est solution duale} \\ p = d \end{cases}$$

~~Preuve~~ Preuve  $\Rightarrow$  Soit  $(x, \varphi)$  point-selle

### Théorème

$(x, (\lambda, v))$  est un point-selle de  $\mathcal{L}$  ??

Condition  
KKT

- $\left. \begin{array}{l} \textcircled{1} \quad x \in \arg \min_{x'} f(x') + \langle \lambda, Ax - b \rangle + \langle v, g(x) \rangle \\ \textcircled{2} \quad Ax = b \text{ et } g(x) \leq 0. \\ \textcircled{3} \quad v \geq 0. \\ \textcircled{4} \quad \text{condition de complémentarité} \end{array} \right\}$   
 $\forall i \in 1 \dots p. \quad v_i g_i(x) = 0.$

Preuve : soit  $(x, (\lambda, v))$  un point selle

$x$  est primal optimal. Donc  $x$  est un point faisable  
 $g(x) \leq 0$  et  $Ax = b$   $\textcircled{2}$  est vraie

$\varphi = (\lambda, v)$  est duale optimal

~~et~~

$x \in \arg \min_x \mathcal{L}(\cdot, (\lambda, v))$

$x \in \arg \min_x (f(x') + \langle \lambda, Ax - b \rangle + \langle v, g(x) \rangle - \cancel{\frac{1}{2} \|x - x'\|^2})$

Donc  $\textcircled{1}$  est vraie

Par ailleurs,

$(\lambda, v) \in \arg \max_{(\lambda, v)} \mathcal{L}(x, \cdot)$

$(\lambda, v)$  maximise  $f(x') + \langle \lambda, Ax - b \rangle + \langle v, g(x) \rangle$

$v$  maximise  $\underbrace{\langle v, g(x) \rangle}_{\sum_{i=1}^p v_i g_i(x)}$  avec  $v \geq 0$

$v_i$  maximise  $v_i g_i(x)$  avec  $g_i(x) \leq 0$   $v_i \geq 0$ .

Si  $g_i(x) = 0$ , pas de contrainte sur  $v_i$  excepté  $v_i \geq 0$

Si  $g_i(x) < 0$ , maximum atteint par  $v_i = 0$ .

On a bien  $v_i g_i(x) = 0$ .

Exercice: Waterfilling

•  $n$  canaux parallèle

•  $P \geq 0$  une puissance disponible à l'émetteur

contrainte contrainte:  $\sum_{i=1}^n x_i \leq P$

Chaque canal a un gain  $a_i > 0$ .

Le débit (capacité) du canal  $i$  est

$$\log(1 + a_i x_i)$$

But minimiser  $- \sum_i \log(1 + a_i x_i)$

$$\vec{x}: \sum_{i=1}^n x_i \leq P, x_i \geq 0.$$

$$x_i \leq 0.$$

$n+1$  contraintes

Si  $x_i = 0$ , alors par ④  $v_i = -a_i + v_0$

Si  $x_i > 0$ , alors  $v_i = 0$  donc

$$0 = \frac{a_i}{1 + a_i x_i} + v_0, v_0 = \frac{1}{a_i^{-1} + x_i}$$

$$x_i = \frac{1}{v_0} - a_i^{-1}$$

Conclusion  $v_0$  paramètre à fixer

•  $v_i + q, v_0 \geq a_i$  on a  $x_i = 0$ .

$$v_i + q, v_0 < a_i, \text{ on a } x_i = \frac{1}{v_0} - \frac{1}{a_i}$$

Finalement, la solution s'exprime

$$v_i, x_i = \max\left(0, \frac{1}{v_0} - \frac{1}{a_i}\right)$$

Reste à trouver  $v_0$ .

## Recherche linéaire

### Méthode du gradient.

$$x_{k+1} = x_k - \gamma Df(x_k) \quad \text{et} \quad \gamma < \frac{\lambda}{L(Df)} \leq \text{constante de Lipschitz de } Df$$

TP: le gradient n'est pas lipschitzien. Par contre, pour tout ensemble compact, si  $f$  est  $C^1$ ,  $Df$  est lipschitzien sur ce compact.

Même si  $L(Df) < \infty$ , elle peut être difficile à calculer et très grande.

Solution: recherche linéaire  $x_{k+1} = x_k - t_k Df(x_k)$ ,  $t_k$  est choisi au cours de l'algorithme.

Petit rappel:

Pour démontrer la vitesse de convergence de la méthode du gradient, on utilise l'inégalité de Taylor Lagrange  $f(x_{k+1}) \leq f(x_k) + \langle Df(x_k), x_{k+1} - x_k \rangle + \frac{L(Df)}{2} \|x_{k+1} - x_k\|^2$

Idée: chercher  $t_k$  qui vérifie

$$f(x^+) \leq f(x_k) + \langle Df(x_k), x^+ - x_k \rangle + \frac{L(Df)}{2} \|x^+ - x_k\|^2 \quad \text{où } x^+ = x_k - \frac{1}{L(Df)} Df(x_k)$$

$$\text{On teste } L_k^s = a_k \geq 0$$

Si l'inégalité n'est pas vérifiée, on multiplie  $L_k^{s+1} = L_k^s \times b$  avec  $b > 1$

dans que la recherche linéaire termine si  $L(Df) < \infty$

On sait que  $f(x^+) \leq f(x_k) + \langle Df(x_k), x^+ - x_k \rangle + \frac{L(Df)}{2} \|x^+ - x_k\|^2$

Donc si  $L_k^s \geq L(Df)$ , alors le test est vérifié et la recherche linéaire s'arrête.

$$L_k^s = a_k^s < b L(Df) \quad \text{donc} \quad b^s < \frac{b L(Df)}{L_k^s} \quad \text{et} \quad s \leq \frac{\ln(b L(Df))}{\ln(b)}$$

(proposition 3.11)

$$\text{Choisir de } a_k \text{ et } b: \quad a_k = \frac{L_k^s}{b}, \quad b = 2$$

méthode du gradient avec recherche linéaire:  $a_k, b_k, b$

$$\text{pour } k \in \mathbb{N}, \quad L_k^s = \frac{L_{k-1}}{b}$$

$$\text{faire: } x^+ = x_k - \frac{1}{L_k^s} Df(x_k)$$

$$t_k = L_k^s b$$

$$\text{tant que: } f(x^+) \geq f(x_k) - c \cdot \langle Df(x_k), x^+ - x_k \rangle + \frac{L_k^s}{2} \|x^+ - x_k\|^2$$

$$x_{k+1} = x^+$$

$$L_{k+1}^s = L_k^s$$

Théorème: l'algorithme vérifie

$$f(x_k) \leq \frac{L(\eta_k) \times b}{2k} \|x_0 - x_k\|^2 + \text{argmin } f$$

dans:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2} \|x_{k+1} - x_k\|^2$$

car  $x_{k+1}$  vérifie le test

$$\forall x, \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2} \|x_{k+1} - x_k\|^2 = \langle \nabla f(x_k), x - x_k \rangle + \frac{L_k}{2} \|x - x_k\|^2$$

donc

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L_k}{2} \|x - x_k\|^2 - \frac{L_k}{2} \|x - x_{k+1}\|^2$$

$$\text{En particulier avec } x = x_k: f(x_{k+1}) \leq f(x_k) - \frac{L_k}{2} \|x_{k+1} - x_k\|^2$$

on prend  $x = x^*$  et on divise par  $L_k$

$$\begin{aligned} \frac{1}{L_k} f(x_{k+1}) &\leq \frac{1}{L_k} (f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle) + \frac{1}{2} \|x_k - x_{k+1}\|^2 - \frac{1}{2} \|x_k - x^*\|^2 \\ &\leq \frac{1}{L_k} f(x_k) + \frac{1}{2} \|x_k - x_{k+1}\|^2 - \frac{1}{2} \|x_k - x^*\|^2 \end{aligned} \quad ???$$

je connais pour  $k \in \{0, \dots, k+1\}$

$$\sum_{k=0}^{k+1} \frac{1}{L_k} [f(x_{k+1}) - f(x_k)] \leq \frac{1}{2} \|x_k - x_0\|^2 - \frac{1}{2} \|x_k - x_{k+1}\|^2$$

$\forall k \leq k+1, f(x_{k+1}) \geq f(x_k)$  et  $L_k \leq L(\eta_k) \times b$

$$\frac{1}{bL(\eta_k)} \times k (f(x_k) - f(x_{k+1})) \leq \frac{1}{2} \|x_k - x_{k+1}\|^2$$

$$f(x_k) - f(x_{k+1}) \leq \frac{bL(\eta_k)}{2} \|x_k - x_{k+1}\|^2$$

$F = f + g$      $f$  dérivable,     $g$  continu

Méthode de Newton :

$$\min f(x) \quad f \in C^2 \text{ avec } \nabla^2 f(x) \text{ inversible}$$

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

Par exemple si  $f$  est strictement convexe,  $\nabla^2 f(x)$  inversible  $\forall x$

Théorème: Si  $f$  est  $C^3$  et  $\nabla^2 f(x^*) \succcurlyeq 0$  alors  $\exists M > 0$  et  $C > 0$  telle que

$\forall \|x_k - x^*\| \leq M$ , alors  $\|x_{k+1} - x_k\| \leq C \|x_k - x_{k+1}\|^2$

$$\text{dans: } x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

Comme  $f$  est  $C^3$ ,  $\nabla f$  est  $C^2$  et donc  $0 = \nabla f(x_k) = Df(x_k) + \nabla^2 f(x_k)(x_k - x_{k+1}) + g(x_k)(x_k - x_{k+1})^2$

$$\text{et } \exists M_0 \text{ tel que } \|g(x_k)\| \leq M_0. \quad x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

$$= x_k - (\nabla^2 f(x_k))^{-1} (-\nabla f(x_k) \cdot x_k - g(x_k) \cdot x_k - x_k)^T$$

$$x_{k+1} = x_k + (x_k - x_k) + (\nabla^2 f(x_k))^{-1} g(x_k) x_k \|x_k - x_k\|^2$$

$$\|x_k - x_{k+1}\| = \|(\nabla^2 f(x_k))^{-1} g(x_k)\| \|x_k - x_k\|^2$$

( $\rightarrow \| \nabla^2 f(x_k) \|^2$ ) et continue

donc pour  $M=1$ , elle est formée sur la boule de centre  $x_0$  et de rayon 1 avec  $C = \max_{x \in B(x_0, 1)} \|(\nabla^2 f(x))^{-1}\| M_0$ , on a bien  $\|x_k - x_{k+1}\| \leq C \|x_k - x_k\|^2$

1.17

## Méthode dual

$$\min f(x)$$

sous contraintes  $Ax=b$   $A: \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,  $b \in \mathbb{R}^m$   $f$  convexe

$$L(x, y) = f(x) + \langle cy, Ax-b \rangle \quad \text{(lagrangien)}$$

$$d(y) = \inf_x L(x, y) \quad \text{(fonction dual)}$$

$$\max_y d(y) \quad \text{(problème dual)}$$

## Méthode des multiplicateurs de Lagrange :

méthode du gradient sur la fonction dualle

J'introduis la transformée de Fenchel de  $f$ :

$$f^*: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$$

$$f^* \mapsto \sup_x \{ f(x) - f^*(x) \}$$

$$\begin{aligned} d(y) &= \inf_x f(x) + \langle cy, Ax-b \rangle = -\sup_x [-f(x) + \langle cy, -Ax+b \rangle] \\ &= -\langle cy, b \rangle - \sup_x \{ \langle -A^T y, x \rangle - f(x) \} \\ &= -\langle cy, b \rangle - f^*(-A^T y) \end{aligned}$$

Si  $f^*$  est dérivable alors  $d$  est dérivable.

lemme : Si  $f$  est  $\mu$ -faisceau convexe, alors  $f^*$  est dérivable et  $\nabla f^*$  est  $\frac{1}{\mu}$ -lipschitzien

$f$   $\mu$ -faisceau convexe  $\Rightarrow f - \frac{\mu}{2} \|x\|^2$  est convexe

$$\text{prop 2.1.4 } \forall g \in \mathbb{R}^n, \forall x \in \mathbb{R}^n, \forall t \in [0, 1] \quad f(tx + (1-t)y) \leq tf(x) + (1-t)y - \frac{\mu(1-t)}{2} \|x-y\|^2$$

$\exists \bar{y} \in \mathbb{R}^n$ ,  $\theta \in \mathbb{R}^n$ ,  $\forall y \in f(x)$        $f(y) \geq f(x) + c\theta^\top y - \frac{\mu}{2} \|y - \bar{y}\|^2$   
 On s'intéresse à  $\frac{f^*(x+th) - f^*(x)}{t}$  pour  $\theta \in \mathbb{R}^n$ ,  $h \in \mathbb{R}^n$  et  $t \neq 0$   
 $\lim_{t \rightarrow 0} \frac{f^*(x+th) - f^*(x)}{t} = c\theta^\top \bar{y}$ ,  $h \rightarrow$  si la limite existe

$$f^*(x) = \sup_{\bar{y}} \langle x, \bar{y} \rangle - f(\bar{y}) \quad \text{lower semi-continuity}$$

$$\forall \bar{y} \in \mathbb{R}^n \quad f^*(\bar{y}) = \langle x, \bar{y} \rangle - f(\bar{y}) \quad \text{car } (\bar{y} \mapsto -\langle x, \bar{y} \rangle + f(\bar{y})) \text{ est scai et croissante}$$

$$f^*(x+th) = \sup_{\bar{y}} \langle x+th, \bar{y} \rangle - f(\bar{y}) \geq \langle x+th, \bar{y} \rangle - f(\bar{y})$$

$$\frac{f^*(x+th) - f^*(x)}{t} \geq \frac{\langle x+th, \bar{y} \rangle - \langle x, \bar{y} \rangle - \langle h, \bar{y} \rangle}{t} \geq \langle h, \bar{y} \rangle$$

Cela montre que  $\bar{y} \in \partial f^*(x)$ , pour l'inégalité dans l'autre sens  
 $\langle x+th, \bar{y} \rangle - f(\bar{y}) \leq ?$

$$\bar{y} = \operatorname{argmin}_{\bar{y}} f(\bar{y}) - c\langle x, \bar{y} \rangle \quad \text{donc } 0 \in \partial f(\bar{y}) \rightarrow$$

$$\langle x+th, \bar{y} \rangle - f(\bar{y}) = \langle x, \bar{y} \rangle - f(\bar{y}) + t\langle h, \bar{y} \rangle \leq \langle x, \bar{y} \rangle - f(\bar{y}) - c\langle x, \bar{y} \rangle + t\langle h, \bar{y} \rangle$$

$$- \frac{\mu}{2} \|\bar{y} - \bar{x}\|^2 + t\langle h, \bar{y} \rangle$$

car  $(\bar{y} \mapsto f(\bar{y}) - c\langle x, \bar{y} \rangle)$  est  $\mu$ -fortement convexe.  $\sup_{\bar{y}} \langle x+th, \bar{y} \rangle - f(\bar{y}) \leq \langle x, \bar{y} \rangle + th$

$$\begin{aligned} \sup_{\bar{y}} & \frac{\mu}{2} \|\bar{y} - \bar{x}\|^2 + t\langle h, \bar{y} \rangle \\ &= -\frac{\mu}{2} \sum_i \frac{\mu}{\lambda_i} h_i^2 + t \sum_i h_i \bar{x}_i \\ &= t\langle h, \bar{y} \rangle + \frac{t^2}{\sum_i} \|h\|^2 \end{aligned}$$

$$\frac{t}{2} (f^*(x+th) - f^*(x)) \leq \frac{t^2}{\sum_i} \|h\|^2 + \langle h, \bar{y} \rangle$$

Par ailleurs,  $\frac{t}{2} (f^*(x+th) - f^*(x)) \geq \langle h, \bar{y} \rangle$

$$\text{Donc } \lim_{t \rightarrow 0} \frac{t}{2} (f^*(x+th) - f^*(x)) = \langle h, \bar{y} \rangle$$

Cela montre que  $Df^*(x) = \bar{y} = \operatorname{argmax}_{\bar{y}} \langle x, \bar{y} \rangle - f(\bar{y})$

$$\|Df^*(x_1) - Df^*(x_2)\| \leq \|\bar{y}_1 - \bar{y}_2\|$$

On a  $\langle x_1, \bar{y}_1 \rangle - f(\bar{y}_1) \leq \langle x_2, \bar{y}_2 \rangle - f(\bar{y}_2) - \frac{\mu}{2} \|\bar{y}_1 - \bar{y}_2\|^2 + \langle x_2 - x_1, \bar{y}_2 \rangle$

$$\frac{\mu}{2} \|\bar{y}_1 - \bar{y}_2\|^2 \leq f^*(x_2) - f^*(x_1) - \langle \bar{y}_2, x_1 - x_2 \rangle$$

on a aussi l'inverse (les rôles),

$$\frac{\mu}{2} \|\bar{y}_2 - \bar{y}_1\|^2 \leq f^*(x_1) - f^*(x_2) - \langle \bar{y}_1, x_1 - x_2 \rangle$$

En combinant les 2 inégalités:  $\frac{\mu}{2} \|\bar{y}_1 - \bar{y}_2\|^2 \leq \langle \bar{y}_1 - \bar{y}_2, x_1 - x_2 \rangle \leq \|\bar{y}_1 - \bar{y}_2\| \cdot \|x_1 - x_2\|$

$$\text{Ainsi: } \|\bar{f}_i - f_i\| \leq \frac{1}{M} \|x_i - x^*\|, \quad \|\nabla f^*(x_1) - \nabla f^*(x_2)\| \leq \frac{1}{M} \|x_1 - x_2\|$$

$$d(y) = -f^*(y - A^T y) - C_b(y)$$

Si  $f$  est  $\mu$ -fortement convexe,  $d$  est aussi dérivable et  $\nabla d$  est  $\frac{\mu M}{M}$ - lipschitzien

Méthode des multiplicateurs de Lagrange:

$$\lambda^{k+1} = \lambda^k + \gamma \nabla d(\lambda^k)$$

$$\text{avec } \gamma < \frac{2M}{\mu M^2}$$

$$\text{De plus, } \nabla d(\lambda^k) = A^T x^{k+1} - b \text{ où } x^{k+1} = \operatorname{argmin}_x f(x) + C_A^T(\lambda^k, x)$$

$$\begin{cases} x^{k+1} = \operatorname{argmin}_x f(x) + C_A^T(\lambda^k, Ax - b) \\ \lambda^{k+1} = \lambda^k + \gamma(Ax^{k+1} - b) \end{cases} \text{ où } \gamma < \frac{2M}{\mu M^2}$$

théorème: si  $f$  est  $\mu$ -fortement convexe

la méthode des multiplicateurs de Lagrange converge vers un point selle du lagrangien

On remplace une minimisation avec contrainte par une succession de minimisations sans contraintes

$$\text{Supposons qu'on veuille résoudre } \min f(x) + g(z) \quad z = Ax$$

$$\text{La MUL devient } \begin{cases} (x^{k+1}, z^{k+1}) = \operatorname{argmin}_{\substack{x, z \\ Ax = z}} f(x) + g(z) + \langle \lambda^k, z - Ax \rangle \\ \lambda^{k+1} = \lambda^k + \gamma(C_{Ax^k} - Mz^{k+1}) \end{cases}$$

En utilisant la séparabilité:

$$\begin{cases} x^{k+1} = \operatorname{argmin}_x f(x) - \langle \lambda^k, Ax \rangle \\ z^{k+1} = \operatorname{argmin}_z g(z) + \langle \lambda^k, z \rangle \\ \lambda^{k+1} = \lambda^k + \gamma(C_{Ax^k} - Mz^{k+1}) \end{cases}$$

Méthode du Lagrangien augmenté:  $\min_x f(x), Ax = b$

Si  $f$  n'est pas fortement convexe,  $d$  pourrait ne pas être dérivable

Mais  $d$  est concave donc on peut appliquer la méthode du point proximal sur  $-d$   
 $\lambda^{k+1} = \operatorname{prox}_d(\lambda^k)$

théorème: Soit  $\lambda^k \in \mathbb{R}^m$

$$\exists x^{k+1} \in \operatorname{argmin}_x f(x) + \langle \lambda^k, Ax - b \rangle + \frac{\gamma}{2} \|Ax - b\|^2$$

$$\lambda_{k+1} = \text{prox}_{\gamma f}(\lambda_k) = \lambda_k + \gamma(Ax_{k+1} - b)$$

La quantité  $f(x) + \gamma \|Ax - b\|^2$  est appelée le lagrangien augmenté